DOCUMENT RESUME

ED 055 111 TM 000 842

AUTHOR Klein, Stephen P.; And Others

TITLE Procedures for Needs-Assessment Evaluation: A Symposium.

INSTITUTION California Univ., Los Angeles. Center for the Study of Evaluation.

SPONS AGENCY Office of Education (DHEW), Washington, D.C. Cooperative Research Program.

PUB DATE May 71

NOTE 63p.; From symposium delivered at the Annual Meeting of the American Educational Research Association, New York, New York, February 1971

EDRS PRICE MF-$0.65 HC-$3.29

DESCRIPTORS *Decision Making; *Educational Needs; Educational Objectives; Elementary Schools; Evaluation Criteria; *Evaluation Techniques; Models; Norms; Principals; *Resource Allocations; Symposia; Test Reliability; *Test Selection; Test Validity

IDENTIFIERS *CSE Elementary School Evaluation KIT

ABSTRACT

Symposium topics and speakers include "Choosing Needs for Needs Assessment" (Stephen P. Klein); "Selecting Tests to Assess the Needs" (Ralph Hoepfner); "Making Better Decisions on Assessed Needs: Differentiated School Norms" (Paul A. Bradley and Dale Woolley); and "Allocating Resources by Subject Area" (James S. Dyer and Guy P. Strickland). A list of 145 goals of elementary education from the CSE Elementary School Evaluation KIT is appended. (AG)

ED055111

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

CENTER FOR THE
STUDY OF
EVALUATION

UCLA
Graduate School
of Education
Los Angeles, California

PROCEDURES FOR NEEDS-ASSESSMENT EVALUATION: A SYMPOSIUM

Stephen P. Klein
Ralph Hoepfner
Paul A. Bradley & Dale Woolley
James S. Dyer & Guy P. Strickland

CSE Report No. 67
May 1971

TM 000 842

CENTER FOR THE STUDY OF EVALUATION

Marvin C. Alkin, Director

UCLA Graduate School of Education

The CENTER FOR THE STUDY OF EVALUATION is one of nine centers for educational research and development sponsored by the United States Department of Health, Education and Welfare, Office of Education. The research and development reported herein was performed pursuant to a contract with the U.S.O.E. under the provisions of the Cooperative Research Program.

Established at UCLA in June, 1966, CSE is devoted exclusively to finding new theories and methods of analyzing educational systems and programs and gauging their effects.

The Center serves its unique functions with an inter-disciplinary staff whose specialties combine for a broad, versatile approach to the complex problems of evaluation. Study projects are conducted in three major program areas: Evaluation of Instructional Programs, Evaluation of Educational Systems, and Evaluation Theory and Methodology.

This publication is one of many produced by the Center toward its goals. Information on CSE and its publications may be obtained by writing:

Office of Dissemination
Center for the Study of Evaluation
UCLA Graduate School of Education
Los Angeles, California 90024

PROCEDURES FOR NEEDS-ASSESSMENT EVALUATION: A SYMPOSIUM*

by

Stephen P. Klein
Ralph Hoepfner
Paul A. Bradley & Dale Woolley
James S. Dyer & Guy P. Strickland

CSE Report No. 67
May 1971

Center for the Study of Evaluation
UCLA Graduate School of Education
Los Angeles, California

*This report is based on a symposium delivered at the American Educational Research Association Annual Convention, New York. February, 1971.

TABLE OF CONTENTS

Page

Choosing Needs for Needs Assessment
By Stephen P. Klein..... 1

Selecting Tests to Assess the Needs
By Ralph Hoepfner..... 9

Making Better Decisions on Assessed Needs:
Differentiated School Norms
By Paul A. Bradley and Dale Woolley..... 20

Allocating Resources by Subject Area
By James S. Dyer and Guy P. Strickland..... 33

References..... 52

Appendix: List of 145 Goals of Elementary Education from the CSE Elementary School Evaluation KIT

CHOOSING NEEDS FOR NEEDS ASSESSMENT

There are three major reasons for determining educational needs. The first of these is to ascertain which needs have the highest priority. Since this information helps to focus the attention of the program planners on the salient problems, it can be used to facilitate planning decisions regarding the modification and development of educational programs. Needs assessment data can thus be used to ensure more efficient utilization and allocation of personnel time and resources. The second reason for conducting a needs assessment is that it justifies focusing attention on some needs and not others. Such justification must often be made in proposals and in reports to school boards and parents. Finally, needs assessment data provides valuable baseline information against which to assess subsequent changes in student performance.

The scope and focus of a needs assessment are, of course, determined by the purposes for which the data will be used. For example, in a school district a needs assessment might be conducted to determine what goals the district should focus on in developing new programs. At a different level, of a particular school, for instance, a needs assessment might be conducted to determine what objectives its ninth-grade mathematics program should try to achieve. Although there are a number of important procedural differences between conducting needs assessments at different levels in the educational system, all needs assessments should include four basic activities:

1. Listing the full range of possible goals (or objectives) that might be involved in the needs assessment.

2. Determining the relative importance of the goals (or objectives).
3. Assessing the degree to which the important goals (or objectives) are being achieved by the program (i.e., identifying discrepancies between desired and actual performance).
4. Determining which of the discrepancies between present and desired performance are the ones most important to correct.

The papers presented in this symposium will discuss these four components. This paper will discuss the first two: listing the range of goals that could be involved in needs assessment and determining their relative importance.

Preparing Sets of Goals

Many of us have experienced the confusion, frustration, and arguments associated with trying to construct educational goals and objectives. We have also learned that it is better to construct these goals in cooperation with parents, teachers, students, and others, so that in the end the goals are more readily accepted. The inclusion of such groups, however, almost always seems to increase the frustration and conflict associated with the goal construction process. In fact, it often seems that by the time the goals have been constructed, the energy and rapport that might have been directed at constructing programs to meet these goals has already been spent. This situation has led the Center to suggest a somewhat different approach to goal selection. The first step of this approach is to have a team of experts construct a set of the full range of goals and objectives that might be included in a needs assessment. Once this set is prepared, however, all the people who ought to be involved in goal selection should be asked to participate. In other words, the strategy we are proposing

is to have experts construct the full range of potential goals that might be included, and then to achieve community, student, and teacher involvement by having these groups participate in the selection of the goals which will be examined in the needs assessment. The total list of goals is not, therefore, a prescription as to which goals the school should try to achieve; rather, it is a varied bill of fare from which the appropriate judges can pick and choose the goals they feel are the most relevant. Thus, the emphasis is placed upon schools selecting which goals they wish to assess rather than on trying to construct just those in which they are most interested.

This approach of using experts to develop the full range of goals also speeds up the construction process. It does this by eliminating many arguments about which goals should or should not be included in a given school's program since the construction process is now limited to describing what might be accomplished by any school as opposed to what should be achieved by a particular school.

If only one school in the country wanted to conduct a needs assessment, this procedure might not be very efficient. In reality, however, almost all schools conduct needs assessments. Thus, it appears that unless this approach is adopted schools will continue to spend considerable time, energy, and funds only to reinvent slightly different wheels.

The Center's Elementary School Evaluation KIT (Hoepfner, Klein, & Bradley, 1970) is a working example of this goal selection approach. This KIT contains a comprehensive set of 106 goals to help the principle select those he wishes to assess. (These goal areas are listed in the Appendix.) The rationale for developing this set was as follows:

(1) it is a waste of valuable time and resources for every principal to review the relevant literature and write his own set of goals when goals overlap so much between schools; and (2) a single, comprehensive set facilitates determining the utility of potential evaluation measures as well as interpreting the data they provide. The set used in the KIT was compiled from a wide variety of sources, including curriculum guides from different parts of the country, recently published elementary school textbooks, national and statewide evaluation studies, basic research studies of psychologists and educators, and reports of various research centers and laboratories. As one might anticipate, these important sources use different classification systems. The goals of the KIT, therefore, are not presented in terms of a single theoretical position, but are organized to permit continued revision and expansion when additional goals are needed.

In constructing the total set of goals for elementary schools, we were forced to compromise with regard to the specificity of the stated goals. Very specific, operationally stated behavioral objectives have the advantage of being easily understood, defined, and measured. Many feel, however, that they tend to be so specific as to limit their usefulness. Further, maintaining comprehensiveness at such a level of precision would result in such an unwieldly list that even to read through it (let alone make relative value judgments) would be a discouraging prospect and totally unrealistic for regular use in the field.

On the other hand, very generally stated goals are often equivocal or vague; they lead quite easily to different interpretations among different individuals. It is difficult, therefore, to find measurement

instruments to match them precisely. For example, no one would deny that students "should have the abilities and skills necessary to engage in the process of science," but this is a rather useless statement unless it is further defined.

These considerations led us to adopt a set of 106 goals at a level of specificity between behavioral objectives and vague intentions. We then printed these goals and a brief description of them on cards, samples of which appear in Figure 1 on the next page.

Goal Selection

Once a comprehensive set of goals has been constructed or obtained from some other source, the next step is to select those which are most relevant to the particular school or program. In other words, the total set of goals is not a dogmatic prescription as to what a school should try to achieve, but rather a varied listing from which to select those of primary concern.

"Who should be involved in this selection process?" is, of course, a critical question. As noted earlier, there is general agreement that involving more people in making a decision will enhance the likelihood of its acceptance. This involvement must be structured, however, in a way that does not inhibit arriving at a final decision. For example, although some schools might be successful in holding something like a town meeting to select goals, others would find this approach too unwieldy.

This situation has led the Center to suggest that schools use packs of goal cards (perhaps printed on IBM cards for ease in subsequent data processing) and have each person involved in the selection process go

Figure 1: Sample Goal Cards

RESEARCH SKILLS

IN SOCIAL STUDIES

Uses reference materials, maps, globes, and encyclopedias. Uses the library, reading, writing, and problem-solving skills to research and write reports on social studies topics, issues, problems, current events, points of view, etc.

SOCIALIZATION - REBELLIOUSNESS

Has a healthy balance between conformity, acceptance, obedience, rigidity, and non-conformity, criticism, and disrespect. Is open-minded and tolerant to new ideas, non-conformity in others. Respects public and private property, shares, cooperates, is respectful and courteous.

DANCE (RHYTHMIC RESPONSE)

Has poise, muscular control, coordination, and rhythm. Responds to the mood, beat, and rhythm of a selection through movement. Expresses himself freely through movement. Learns popular and folk dances.

INDEPENDENT APPLICATION

OF WRITING SKILLS

Appreciates the importance of good grammar to clear communication. Appreciates writing as a means of self-expression, as a creative endeavor, and as an important means of communication. Enjoys writing activities. Finds satisfaction in having written something well. Takes pride in turning in neat work.

through his deck and indicate those goals he feels are most important. In the Elementary School Evaluation KIT, this process is facilitated by the provision of five envelopes into which the cards may be sorted. These envelopes are labeled to denote the relative importance of goals. The relative importance ranges from "1. Unimportant, Irrelevant" to "5. Most Important." Since each person rates his own deck of 106 cards by placing each card in the appropriate envelope, one can involve as many people in the selection process as there are decks of cards. In fact, since the decks are reusable, there is really no limit on the number of raters other than that imposed by the available time and resources which can be allocated for collecting and analyzing the data in this step of the needs assessment.

Once the ratings are gathered from all of the people involved in the goal selection process, there are a number of ways of organizing and summarizing the results. These methods range from taking the simple average rating among all the raters to computing weighted averages for subsets of raters, such as parents and teachers. The particular technique chosen will, of course, be a function of the number of raters involved, the political context in which the needs assessment is being done, available time and facilities, etc. Whatever the technique chosen, however, the end result should provide a score for each goal in the whole set. This score indicates that goal's relative priority to the other goals. In short, the scores reflect the value system of the people who rated the goals.

Summary and Conclusions

This paper has presented a new technique for conducting the initial steps of a needs assessment. The essence of this technique is that comprehensive sets of goals or objectives should be constructed by experts

who have the time, knowledge, and resources to fully cover the field of potentially relevant goals and objectives. The second step in this process is to have appropriate individuals select from an appropriate set those goals which are most relevant for their particular situation. In other words, the total set of goals or objectives does not prescribe what a school or program should do, but rather provides a catalog from which selections can be made. The major advantages of this approach are that it can involve many more people than the traditional committee approach of constructing goals and objectives, and it can accomplish this goal selection task quicker and at significantly less cost and frustration.

The other papers presented in this symposium will discuss the actions to be taken following the identification of important goals.

SELECTING TESTS TO ASSESS THE NEEDS

Tests and questionnaires are used to gather evaluation data because they generally are the most efficient means for doing so. They provide more and higher quality information at lower cost than do other assessment techniques. The decision as to which test to use is often a difficult one, however, since existing tests and measures differ widely in the quality and quantity of evaluation information they provide. The problem is often compounded further by misleading claims of test publishers and by complicated technical manuals.

To minimize the difficulty of selecting tests, it first was necessary to have independent test experts evaluate essentially all the existing published tests for elementary school pupils. The four basic criteria used in this analysis were as follows:

1. How well does the test measure the educational goal?
2. To what extent is the test appropriate for the students?
3. To what degree can the test be easily utilized in the school?
4. Is the test sufficiently reliable and refined in measurement?

The complete set of test reviews, organized by grade level and by objective, is presented in CSE Elementary School Test Evaluations (Hoepfner, et al., 1970).

In order to appraise equably the output measures used in elementary schools today (mostly tests of achievement and aptitude), a critical method of test evaluation was developed. Preparatory to the evaluation, all those tests presently available were located and compiled. The tests were then evaluated in order to identify and endorse those most appropriate, effective, and useful in assessing schools or students.

Four evaluation criteria comprise the test evaluation system labeled the "MEAN" method, an acronym for the four criteria:

1. Measurement validity,
2. Examinee appropriateness,
3. Administrative usability and
4. Normed technical excellence.

All the output measures prepared for, or potentially useful for, evaluations within the elementary schools that are generally available to educators and researchers were evaluated on the above four critical assessment criteria. Also, each subscale of a measure was evaluated separately if it had been normed or was recommended for use in decision making. The four criteria comprising the MEAN system will be described in this paper.

Measurement Validity

This criterion is essentially a measure of psychological validity. Empirical measurements of such validities were most desired, but indirect evidence was also taken into account. In addition, evidence for correlative validity was weighted.

Evaluators were trained to use the Center's list of educational goals designed to categorize elementary school outputs meaningfully and exhaustively; each test was then judged according to its capacity to assess the particular goal that was determined most appropriate to it. Decisions as to which goal was most appropriate to a test were not based merely upon the goal implied by the test name or on the stated objectives usually given in the test manual. The evaluators went to the individual items

to determine which goal the plurality of the items reflected. A consensus among the evaluators then determined the educational goal by which the test would be evaluated.

This procedure may, of course, unjustly penalize some otherwise excellent test instruments, particularly those constructed on a model of educational goals that differs substantially from those adopted for the this test evaluation program. It appeared, however, that such situations were not common and, in fact, that very few tests of educational output are based upon any explicit model of education or evaluation. It also appeared that evaluation could not logically proceed on a global level, since concepts like "has developed social skills" must be analyzed and refined into reasonably small units in order for them to have much meaning in any evaluation program.

Examinee Appropriateness

The second evaluation criterion is designed to assess how appropriate the test is for the students who will be taking it. Concern was directed toward the appropriateness of the test's comprehension level(s), the physical format, and the manner in which a student records his answers.

The test's comprehension level included two aspects: content and instructions. Evaluation of the appropriateness of test content centered upon the difficulty of the semantic or numerical items, and also upon the relevance or interest-arousing aspects of the items. Instructions were evaluated on clarity, completeness, and complexity.

The second major area where appropriateness is felt to be important is that of test format. The visual principles employed in a test-page

layout were evaluated in terms of effective use of visual principles and design. The evaluators looked for specific format features such as sufficiency of white space between items, visual coherence of item stems and alternatives, and effective use of color as an aid in separating items.

In addition to the whole-page format, the evaluation considered the quality of illustrations and print. Pictorial and geometric item material was evaluated according to meaningfulness and ease of decoding for young children. Evaluations of print were made on the basis of clarity, size, and type-face, at all times considering the limitations of the examinees.

The psychometric problem of speededness vs. power of a test was also considered in the evaluation of appropriateness. For each scale, pacing or time limits were judged as to their appropriateness for the subject matter and for the examinees. In almost all cases, power (i.e., relatively unhurried conditions) was preferred to speed as an attribute of tests of educational output.

The last aspect of appropriateness considered was the type of response recording. The more simple and direct the connections were between the item stem and the recording of a response, the more credit was given. Complicated conversions from item stems to alternatives to unusual or novel answer sheets were given less credit as being generally too complicated, especially for the lower grades.

Administrative Usability

After asking questions such as "What will it measure?" and "Is it designed for my students", the next logical question should be concerned

with how usable the test is in terms of administration, scoring, interpretation, and decision making. These utilization questions comprise the third evaluation criterion of the MEAN method: Administrative usability.

It was assumed that for general assessment of educational output, a test that can be administered to a large group is desirable. Small-group and individually administered instruments, although having their unique advantages, were judged to be less efficient for educational evaluation. It should be noted that all individually administered tests therefore suffer from this evaluative decision, and consequently their ratings indicate less usability.

A second variable strongly affecting a test's utility is the training necessary to administer the test appropriately. Since few schools have resident psychometrists and district psychometrists generally focus their attentions on individual student problems, a test has more utility if it can be administered by the school staff, preferably the students' teacher. The time necessary for test administration also affects its utility. Under the assumption that the average class "unit" of time in elementary schools is about 40 to 45 minutes, tests were credited if they fit into one such time unit, but were not credited on this aspect if their lengths necessitated special scheduling.

The utility of a test is further affected by its scoring procedure. Simple and objective hand or machine scoring of tests was considered optimal for utility, while difficult and subjective scoring received respectively less credit. Although the general utility of tests is not much altered by slight variations in scoring difficulty, it was

decided that tests scored on a purely subjective basis, i.e., many projective techniques, should not even be considered as reasonable candidates for educational evaluation instruments.

From a pragmatic viewpoint, while ease of administration and scoring are desirable, they are dwarfed by the importance of being able to interpret the scores and then reach a decision. Scores can only be interpreted normatively through some type of score standardization or conversion. If the score conversion is to be a trustworthy one, the procedures must be empirical. The empirical conversions are obtained through normed samples which have been given the test under standard conditions.

The samples used in test norming were evaluated according to two criteria: breadth and representativeness. A broad normative sample is one which ranges over a measured dimension greater than the group to which the test is directed. One could then know about extreme performances, either high or low.

Representativeness of the normative sample is concerned with the procedures used in obtaining the comparison group. While purely local tests can be quite adequate measurement devices, the trend in educational evaluation is not in that direction. With national questions being asked, federal support for education and related research being given, and national problems to be solved, a representative national normative sample becomes a most desirable quality of educational tests. The criteria valued for a normative sample were currency, representation of geographic regions, ages, racial and ethnic origin, population density, and variety of schools and school districts. It might be important to note here that few test publishers have done their normative sampling

very well, and that the technical manuals abound with confusing if not downright misleading sampling techniques.

After the test has been administered to its normative sample, the raw scores from that sample are isomorphically mapped into some standardized score conversion system. The normative score conversions were evaluated according to three criteria. If the derived scale is common and generally understood, the test is given credit. If the conversion to the derived, normed scores is clear, with unambiguous tables presented and described, the test earns credit over those with complicated, multi-stage conversions. These two aspects of the derived scores determine in part who can interpret them. Tests yielding scores interpretable by the school staff were preferred to those demanding the skills of a psychometrist.

The final practical consideration of a test's usefulness was whether or not decisions, either individual or group, can be made. Tests which have manuals describing well both score interpretation and subsequent decisions that might be made were evaluated as better than those that have doubtful decision-making utility. The decisions that were considered ranged from selection of the next textbook for a class to whether or not the child should be referred to a specialist for remedial instruction or psychiatric help.

Normed Technical Excellence

The last major criterion of the MEAN evaluation procedure, Normed technical excellence, is concerned with the reliability, replicability, and refinement of measurement of the tests. The standard approaches to test reliability are not vitally relevant to tests of educational achievement, although the underlying concepts of reliability theory are.

While test-retest reliability, assessing the long-range stability of a measure (and the examinee), is important for long-range prediction, the notion of long-term stability of examinees' achievements is diametrically opposed to the goals of education. The fatalistic concept of relative stability within groups of students, while not in contradiction to educational goals (although perhaps in contradiction to many educational philosophies), is perhaps the most relevant aspect of stability measurement for long-range prediction. In other words, stability measures are betting on no real change in relative scores over time.

Internal-consistency reliability estimates indicate how coherently the test items assess the same dimension(s) of behavior. This type of reliability also has marginal value in the assessment of educational achievement, since the more internally consistent a test is, the more coherent, and therefore similar, are the test items. Typically, however, achievement tests must assess a broad range of specific educational objectives. It is concluded from the technical manuals of tests that most test publishers do feel that internal consistency among test items is desirable. Whether this decision of the publishers rests upon psychometric judgments or the fact that internal-consistency reliability estimates can be inflated easily by numerous extraneous test qualities remains unknown.

A third estimate of test reliability evaluated is the alternate-form type, when alternate forms are available. When instructional treatment effects are studied, alternate forms of a test are particularly desirable.

Since all three types of reliability estimates are more or less relevant to questions of educational achievement to an equal degree, they were all given equal credit as aspects of the MEAN evaluation procedure. This tactic was necessitated by the fact that selection of any one of the estimates with omission of the remaining two would do violence to the fourth-criterion rating for many of the test instruments.

Closely related to the concept of test reliability is that of replicability of procedures to obtain the achievement scores. If procedures described in test manuals are complicated, subjective, or based upon abnormal samples, the test is clearly not replicable in its findings and therefore is less useful for the educator.

The range of coverage is also an important aspect of a test's technical excellence. A restricted range of assessment, i.e., measurement of a narrow band of achievement like the second month of third-grade geography, limits the test's interpretability. A test which is appropriate for one level of assessment but can also be applied to students from one to two years above and below that level has obvious advantages since both advanced and retarded students can be compared with the normative sample.

Related to range of coverage is the refinement of gradation of the inter-individual comparison scores. Tests yielding scores graduated into centiles or grade placements were rated as well graduated; deciles, stanines, and similar scales as poorly graduated or uncommon; pass-fail, quartiles and novel scales as poorly graduated and uncommon.

The primary concerns of applying the MEAN system were the objectivity and consistency of the evaluations. To maximize both the

objectivity for any one test evaluator and the consistency with which several test evaluators would evaluate, specific guidelines for evaluation of each aspect of each criteria and for letter-grade assignment were developed. These appear in Figure 1.

At least two psychometrically trained educational researchers independently evaluated each test or subscale published or normed. Each measure was independently placed into an educational goal category and then rated by the MEAN system. A third and sometimes a fourth trained researcher then adjudicated the goal assignment and the MEAN ratings.

Figure 1

MEAN TEST EVALUATION FORM

Test Name _____ Form _____ Rater _____ Date _____

Evaluation Criteria	Rating (circle one number in each row)						
1. Measurement Validities a. Content and Construct	0 (only in name)	2 (a few)	4 (some)	6 (fair job)	8 (best available)	10 (hit nail on the head)	M Total
b. Concurrent and Predictive	0 (none reported)	1 (very little)	2 (some)	3 (not enough)	4 (considerable)	5 (exhaustive)	Grade
2. Examinee Appropriateness a. Comprehension: content	inappropriate 0	doubtful 1	possibly appropriate 2	probably appropriate 3	exactly right 4		
instructions	0	1	2	3	4		
b. Format 1. Visual principles	0 (complicated)		1 (probably good)		2 (outstanding aids)		
2. Quality of illustrations (print)	0 (not good)		1 (helpful)		2 (excellent)		
3. Time and pacing	0 (bad)			1 (appropriate for broad range)			E Total
c. Recording answers	0 (complicated)		1 (standard)		2 (especially easy)		Grade
3. Administrative Usability a. Administration 1. Test administration	0 (individual)		1 (small groups)		2 (large groups)		
2. Training of administrators	0 (psychometrist)			1 (school staff)			
3. Administration	0 (43+ minutes)			1 (42 minutes or less)			
b. Scoring	0 (subjective)		1 (difficult)		2 (simple)		
c. Interpretation 1. Norms a. Norm range	0 (restricted)			1 (broad)			
b. Score interpretation	0 (uncommon, abstruse)			1 (common, simple)			
c. Score conversion	0 (complicated)		1 (simple)		2 (clear, tables)		
d. Norm groups	0 (local, outdated, or poorly sampled)			1 (national, well sampled)			A Total
d. Score Interpreter	0 (psychometrist)			1 (school staff)			
e. Can Decisions Be Made	0 doubtful		1 possible		2 probable	3 yes – charts and graphs	Grade
4. Normed Technical Excellence a. Stability	not reported or less than .70 0		.70 to .80 1		.80 to .90 2	.90+ 3	
b. Internal Consistency	0		1		2	3	
c. Alternate form	0		1		2	3	
d. Replicability	0			1			
e. Range of Coverage	0 no information		1 floor or ceiling reached		2 adequate	3 more than adequate	N Total
f. Scores	0 poorly graduated and uncommon		1 poorly graduated or uncommon		2 well graduated and standard		Grade

Although the MEAN criteria are relatively complex, within each one of the four evaluative categories a total letter grade was determined to reflect the desired assessment aspects in proportion to their desirability. Points were assigned to each aspect of each criterion in such a way that there would be discrimination for each aspect. The total letter grade, assigned for each major criterion, only indirectly reflects the separate-aspect evaluations.

Each measure earned four letter grades by the MEAN system. The four-letter combination serves as the Center's official evaluation of the test. Should the goals of the user not coincide with those of the CSE Elementary School Evaluation Project, then the MEAN evaluations may be interpreted with different emphasis.

MAKING BETTER DECISIONS ON ASSESSED NEEDS: DIFFERENTIATED SCHOOL NORMS[1]

The procedures in a needs assessment evaluation which have already been described include the selection of educational goals that are to be evaluated and the selection of the best available instruments to assess student performance on these goals. After their selection, the assessment instruments are then administered and scored. The information provided by the assessment devices becomes one of the inputs to the final phase of needs assessment evaluation: the selection of the one or more educational goal areas in which revisions in the instructional program will be made so as to improve student performance.

This last phase in the needs assessment evaluation is the critical one, obviously, as it pinpoints where a school is going to devote some time, effort and, probably, money to correct a deficiency in its instructional program. Making a bad decision at this phase would have dire consequences; expenditures of time, effort, and money in behalf of the selected goal(s) would be wasted, and another goal which better deserved attention would have been neglected. It is imperative, therefore, that a school have the best information possible before it decides which educational goal to select as the target area for improving student performance.

One type of information that is an input to the last phase of a needs assessment evaluation is the data obtained from the assessment of student performance. This paper is concerned with ways in which this

---

[1]An expanded version of this paper, including a review of previous efforts to obtain differentiated school norms, will appear in a forthcoming issue of Evaluation Comment.

information can be improved so that it is maximally useful to the school which is involved in a needs assessment evaluation.

What is the outcome when a school assesses student performance with standardized instruments? First of all, the outcome depends on what the publisher of the instrument makes available. Since all, or nearly all, publishers provide tables of norms, a school could prepare a roster of the raw and scaled scores achieved by every pupil. It is to be understood that from this point on we are talking about student performance within a given grade level. At no time are we looking at or comparing the performance of students in different grades. This notion follows the common practice of interpreting test results relative to the grade level of the student. The most frequently reported scaled scores are centiles, grade equivalents, and stanines. An ambitious person could take this roster and compute averages for each grade level, and if this information were available for other schools in the district then he could compare averages across schools. If a battery of tests had been administered it would then be possible to prepare a profile of achievement for every pupil. However, very infrequently is there a person in an elementary school who either has the time or the experience to undertake such an endeavor.

The larger publishing houses make available several services that aid the school in the interpretation of test results. These services include providing information similar to that described above: that is, rosters of scaled scores, various descriptive statistics for grades, schools, or systems, and individual pupil profiles. In addition, a publisher may indicate the procedures to be followed if a school wished to develop percentile scores for students

either within a school building or within a school system. However, most of the information that can now be provided by test publishers is useful only for evaluating the current status of individual pupils. That is, the various scaled scores that publishers provide indicate the goodness of a student's performance relative to the performance of all students who took the same test.

There are two reasons why the information that is typically available neither is the best information possible nor is maximally useful to a school that is engaged in a needs assessment evaluation. One reason is that virtually all currently available test norms are pupil norms; that is, they indicate the relative goodness of an individual student's raw score. The second reason is that, again, virtually all test norms are national norms based on samples of students that are intended to be representative of all students in the country. Why do these reasons make the typical test norms inappropriate for a needs assessment evaluation?

School Norms

In a needs assessment evaluation, the unit being evaluated is the school, not a single student. One aspect of a needs assessment evaluation is determining how well a school is producing appropriate student achievement in the chosen educational goal areas. That is, once a school has identified the most important educational objectives, it must determine its level of achievement on these educational objectives. It is not possible to determine the school's level of achievement from pupil norms, as these norms are inappropriate. What is needed instead, for the purpose of a needs assessment evaluation, are norms that would give the relative goodness of

the school's performance on a standardized test. Percentile norms could be derived for schools very easily since the process would be the same as that used in deriving pupil norms. The one necessary change is that the school's mean raw scores on the test rather than the pupils' raw scores would be used to compute percentiles.

One might, at this point, wonder why a school cannot determine its level of performance by looking up its mean raw score on a standardized test in a table of pupil percentile norms. It is not appropriate to do this because a school would get an incorrect indication of its level of performance. The difference between pupil and school norms is based on the fact that there is less variation in school means than in pupil raw scores. Figures 1 and 2 illustrate this difference. Figure 1 shows hypothetical normal frequency distributions of pupil raw scores (A) and school scores (B). It is seen that there is less variation in the school scores than in the pupil raw scores.

A normal frequency distribution was chosen for convenience only. No implication is intended that actual pupil or school frequency distributions have the characteristics of a normal distribution. It is also a convenience that the means of the distributions are the same. The standard deviation of the pupil scores is 10 while that of the school scores is 5. No generalization is possible regarding the ratio of standard deviations of pupil scores and school scores other than that the former is larger than the latter. Again, the standard deviations of 10 and 5 were chosen for their graphical and conceptual impact.

Figure 2 shows the cumulative proportions of the frequency distributions in Figure 1. The curves in this figure can be used to read the pupil and school percentile scores. Curve A gives pupil percentile scores; curve B gives school percentile scores. For example, if a pupil's score is 24, then his percentile score is 27. But if a school's score is also 24, then its percentile score is 11. If one looks at a score of 37, however, it is seen that the pupil percentile is 75 and the school percentile is 92. Thus, looking at a raw score that would fall below the 50th percentile, a school's percentile score is lower than a pupil's percentile score, but looking at a raw score that would fall above the 50th percentile, a school's percentile score is higher than a pupil's percentile score.

Differentiated Norms

There is yet another way in which the norm tables provided by most publishers can be less than adequate for needs assessment evaluation. In most instances, for better or worse, the norm tables are based on a national sample of schools. Thus, even if a publisher did produce school percentile norms, a school's performance would be compared to the performance of all schools in the sample. But what if there were certain characteristics of schools, characteristics outside the students' cognitive and affective skills, that were found to be related to their level of performance? The characteristics of a school that are pertinent here are often referred to as input variables. An example of an output variable would be school performance on a standardized achievement test. The input variables could be such things as the number of volumes in the

school library, the average expenditure per student, the occupational level of parents, and the racial mixture of the students. It is not even necessary to hypothesize about this situation; the Coleman study, for example, has shown that such relationships do exist. Under these conditions, then, the use of national norms can lead to an unfair and biased comparision if a school is atypical in its characteristics.

The bias can lead to both over- and underestimation of a school's level of performance. This bias can be easily illustrated through the use of cumulative proportion curves. Suppose it were found that there were three different "types" of schools which had markedly different performance on a standardized test. Figure 3 shows the cumulative proportion curves (A, B, C) for the three hypothetical types of schools as well as a cumulative proportion curve for all the schools (D). The three cumulative proportion curves A, B, and C correspond to normal frequency distributions with means of 20, 30, and 40, respectively. All have a standard deviation of 4. The cumulative proportion curve D corresponds to a normal frequency distribution with mean 30 and standard deviation 8. Again, these means and standard deviations were selected for convenience and graphical impact. No implication is intended that these curves represent distinct possibilities. It is not known how much separation of schools can be achieved, how school scores are distributed, and what the relationship among standard deviations is.

Case 1. Consider the four outcomes indicated in Figure 3. First of all, suppose a school's score was 14. If that school found its percentile rank from curve D on Figure 3, it would find that its performance fell at the 2nd percentile. It is possible that a principal confronted

with this result would immediately begin a litany of alibis to account for his school's low performance. Many of these alibis are likely to refer to the inputs to the school, such as a low SES level, a variable ethnic composition, poor tax support, low teacher salaries, etc. However, suppose that curve A in Figure 3 gives the percentile ranks for schools that are similar to the principal's school in terms of input variables. Using this curve (A), the principal would find that his school's performance fell at the 6th percentile. That is, when compared to schools that have similar resources, his school did better than only 6% of the schools. This result should indicate, rather unequivocally, to the principal that his school is not doing a good job in producing student performance in the area that the test measures.

Case 2. Now consider a school whose score was 21. With reference to curve D, which represents a table of national norms, a score of 21 corresponds to a percentile rank of 13. Again, it is likely that the principal of this school would echo the sentiments of the principal whose school's performance fell at the 2nd percentile. However, when this principal compares his school's score of 21 with similar schools, he learns that his level of performance falls at the 60th percentile, not the 13th percentile. Needless to say, this principal is not likely to be disappointed with such an outcome, and may even be mildly pleased to outrank 60% of the schools.

Case 3. The third outcome indicated in Figure 3 is a score of 38. The percentile rank of this score is 84, with reference to curve D (national norms). A principal who finds that his school falls at the 84th percentile may possibly engage in some strutting and issue

proclamations attesting to the superiority of his school. Suppose, however, that this school is of a type that is characterized by favorable values on the input variables that are related to student performance. For example, some of the characteristics of this type might be high SES level, high teacher salaries, and an all white student body. The percentile ranks for this type of school are given by curve C in Figure 3. Using this curve, a score of 38 corresponds to a percentile rank of 31. A markedly different picture of this school now emerges. Rather than doing a good job with the students, the school appears to be somewhat deficient in producing student output commensurate with its input characteristics.

Case 4. Lastly, consider a school whose score is 46. On the national norms this school falls at the 97th percentile, and the principal of this school would probably be jubilant. Being aware of the quality of the input characteristics of his schools, he may wonder if his school's performance is really as good as it seems. This principal, at least, is in a good position, because his score of 46 falls at the 93rd percentile with reference to curve C.

What are the consequences for needs assessment evaluation when a school uses national norms rather than differentiated school norms? In virtually all cases a school would have an incorrect indication of its relative success. In many of these cases, a school would reach the same decision to select a particular goal area for curriculum revision no matter what norms were used. But in some of these instances, using the national norms rather than differentiated school norms will lead to one of two types of errors: selecting a goal area for

curriculum revision that in fact does not need it, and not selecting a goal area that in fact does need curriculum revision. It is impossible to predict when these errors will occur because there are inputs other than test scores to the decision-making process of selecting a goal area for curriculum revision. (This phase of needs assessment evaluation will be discussed in the last paper in this symposium.)

It should be reiterated at this point that the notion of having different norms for different types of schools is one whose feasibility depends on finding types of schools that differ in their performance on standardized tests. This latter point is important, because while it may be possible to group schools into a small number of categories based on similarities of input characteristics, it may not be the case that there are significant differences in level of student performance. If there is no difference between the groups in level of performance, then there is no need to have three separate norm tables which are essentially identical to each other. It should be remembered, though, that the situation is different with regard to the notion of having tables of school norms as well as tables of pupil norms. This notion is not only very feasible and plausible, it has been done by some publishers.

To summarize, improvements need to be made in the information that results from an assessment of student performance. It is important to improve this information because it is an input to the last phase of a needs assessment evaluation: the selection of the one or more educational goal areas in which revision in the instructional program will be made so as to improve student performance. Specifically, it was proposed that improvements can be made by altering the types of norms that

accompany standardized tests. The two alterations suggested were to provide school norms as well as pupil norms and to provide, if feasible, norms for different "types" of schools as well as national norms.

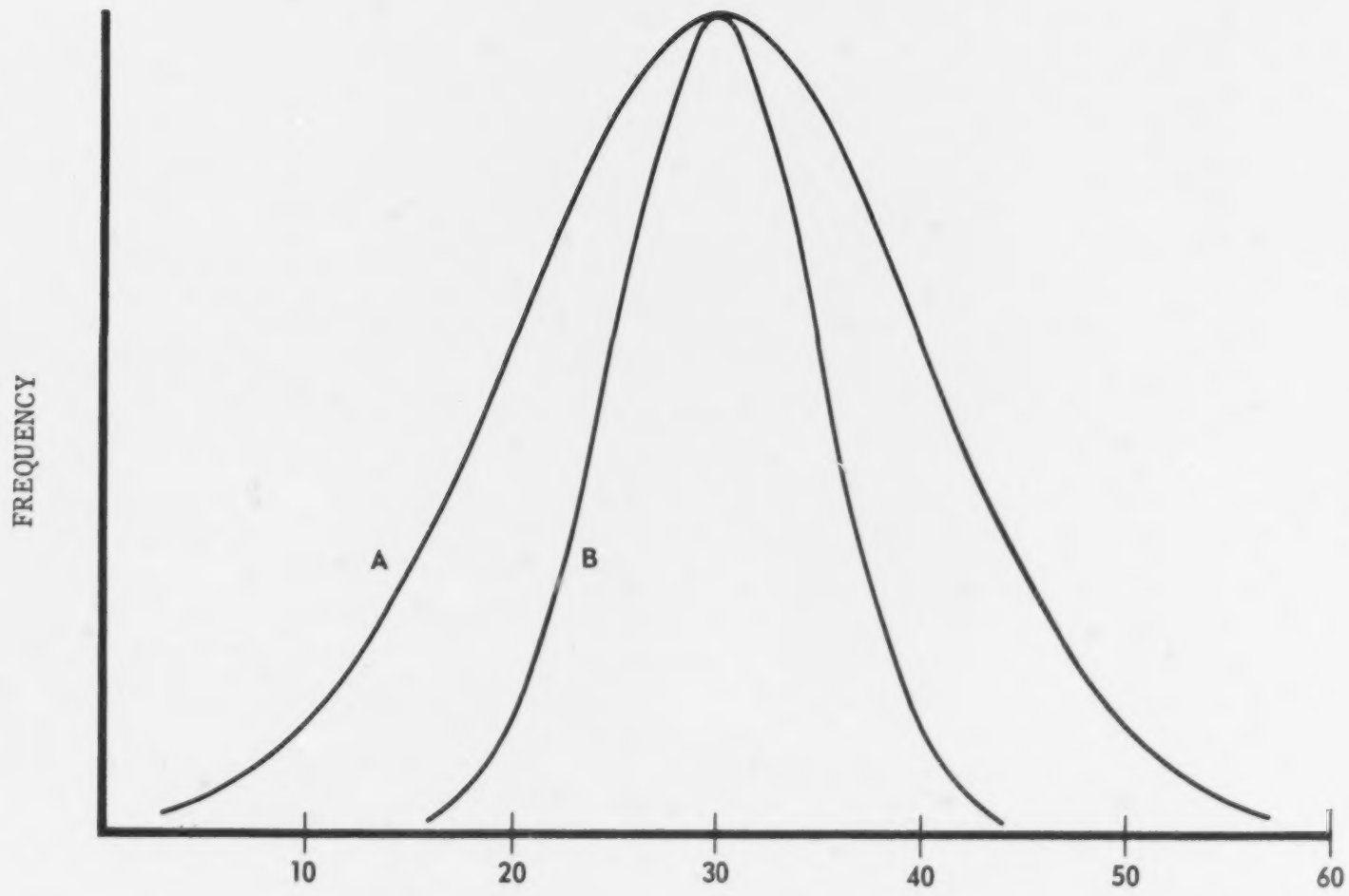

FIGURE 1. HYPOTHETICAL FREQUENCY DISTRIBUTIONS OF PUPIL SCORES (A) AND SCHOOL SCORES (B).

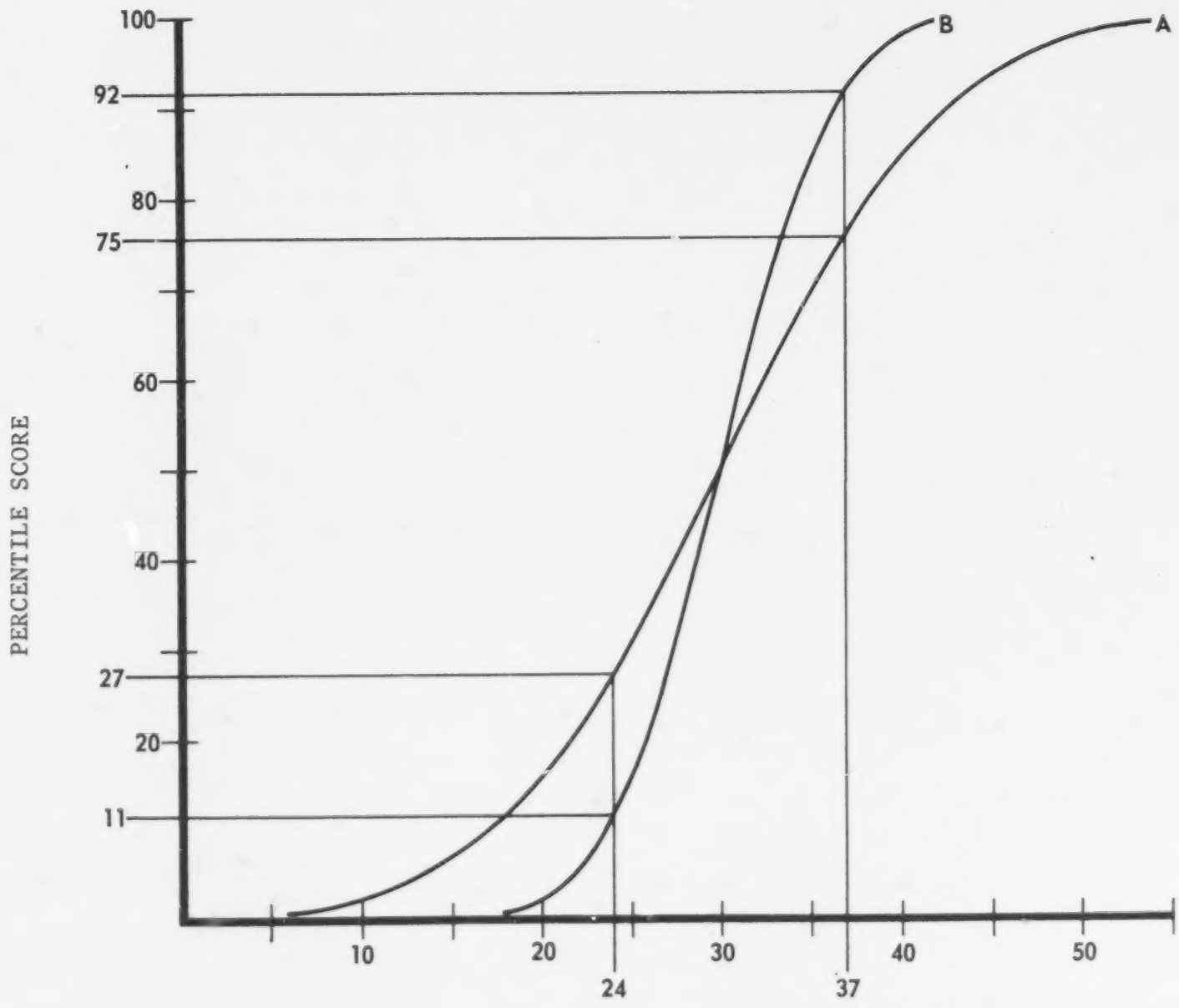

FIGURE 2. HYPOTHETICAL CUMULATIVE PROPORTIONS OF PUPIL SCORES (A) AND SCHOOL SCORES (B).

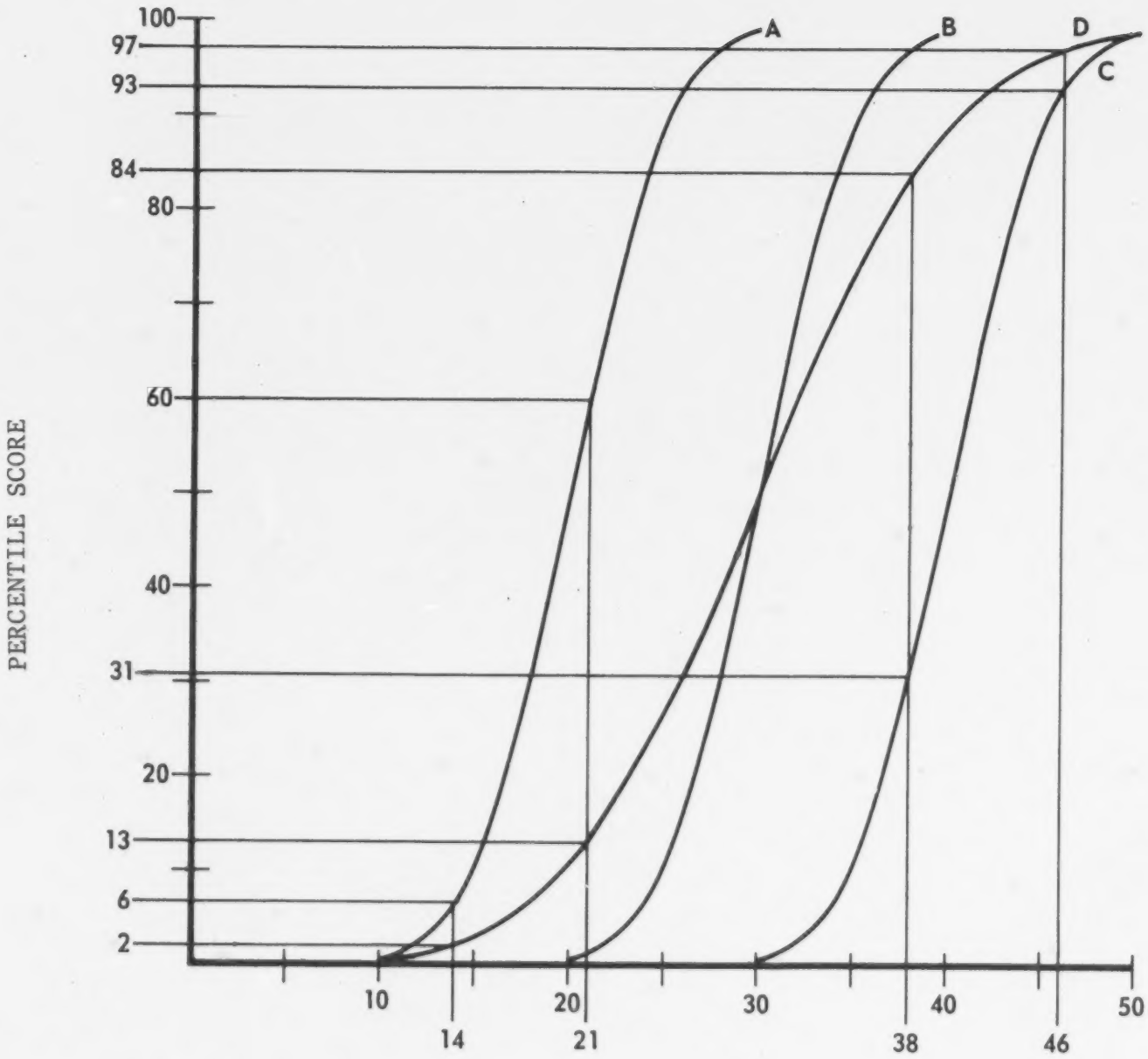

FIGURE 3. HYPOTHETICAL CUMULATIVE PROPORTIONS OF SCHOOL SCORES FOR THREE DIFFERENT TYPES OF SCHOOLS (A, B, C) AND FOR ALL SCHOOLS (D).

ALLOCATING RESOURCES BY SUBJECT AREA

This paper describes a procedure designed to assist elementary school principals in the process of selecting educational subject areas which should command their attention, resources, or support. For each subject area, the model produces an index number which represents the expected "value" which will accrue to the school from the adoption of an instructional program appropriate for strengthening that area. Although the procedure will be explained in terms of this specific application, the approach proposed in this paper could also be used to structure similar decision problems at the district or state levels, or in secondary and pre-school educational systems.

The calculation of the index number for a particular area depends on the following factors: (1) the relative importance of that area; (2) the "utility," or "value" to the decision maker, of making an improvement in that area, given the current level of performance; and (3) the probability distribution of the results of implementing a particular type of improvement program for that area, given the current level of performance. The first two factors will be discussed in the first section of this paper. In the second section the probability of various results will be considered, and all three will be combined in a formula yielding the desired index number. The use of these indices as an aid to decision making will then be explained.

THE ESTIMATION OF UTILITY

This section describes how the "utility"[1] (of the decision maker) for the current state of the system is estimated. In our particular application, the system is an elementary school, the decision maker is its principal, and the state of the system is represented by the level of educational achievement of the school. As we shall see, the process of estimating utility presupposes the existence of (1) a well-defined hierarchy of system objectives, and (2) adequate devices for measuring the degree of achievement of these objectives. We are then left with the problem of transforming performance measurements into a single (utility) number; this number reflects the decision maker's "satisfaction" with the state of the system as represented by these performance measures.

The formal structuring of a decision problem (in our case, the selection of educational subject areas to be emphasized or strengthened) must begin with the statement of a general goal -- perhaps one as vague as "promote the good life". By asking <u>how</u> a given system contributes to the achievement of this "meta-objective," a hierarchy of primary objectives, secondary objectives, goals, and subgoals can be identified.

However, the performance of an educational system is not usually determined by a <u>direct</u> measurement of how well it achieves its primary or secondary objectives. These objectives are too broad in nature for the development of valid and reliable measuring instruments. Instead, the

---

[1]For a discussion of the meaning of "utility," see the classic work by J. Von Neumann and O. Morgenstern (1953) or that of Schlaifer (1959).

state of the system may be assessed indirectly by measuring its performance in each terminal goal area for which adequate measuring devices exist. As discussed earlier in this symposium, the School Evaluation Project has investigated the existence of test instruments for each of the goals listed in the Appendix. The results of this study indicate that numerous instruments are available in the skill and cognitive areas, with their number and validity decreasing in the affective areas. However, continuing improvement in the number and validity of tests in all areas can be expected.

The results achieved on the various test instruments are generally expressed in terms of percentile scores or rankings on a national basis. Consequently, there is no absolute, invariant scale against which to measure performance, and we must take these scores to be our "raw" system performance measurements. It seems reasonable to assume that the "worth" or "value" (to the principal) of a given percentile score depends strongly on both (1) the particular goal area involved and (2) his aspiration level for that area. Since past achievements of a school depend to some extent on such exogenous input factors as the socioeconomic status of the parents, the location of the school (urban vs. rural), the region of the country, etc., one can reasonably expect these factors to influence the principal's aspiration level (and hence his utility for performance measurements) in each goal area.[2] So that the model will be sensitive to

[2]This will be discussed in a forthcoming Center Report on the possible impact of such factors on the shape of the utility function for a given goal area.

these environmental factors, performance data is currently being collected on schools categorized according to such environmental characteristics. (See Elementary School Evaluation KIT: Needs Assessment, Booklet IV.)

If we accept the (percentile) scores obtained on standardized tests as adequate measures of the school's performance in the associated goal areas, it remains to be determined if these scores can be used directly in estimating the principal's utility. The contents of the previous paragraph and the following observations suggest that they cannot; i.e., the scores must be transformed before they can be used for that purpose:

1. Given results in a particular goal area and ignoring all other results, it is clear that a score of 80 may not be considered twice as "good" as a score of 40. Also, it may not be true that a score increase from 40 to 50 has the same "value" as an increase from 80 to 90.
2. The "worth" of an increase from, say, 40 to 50 percentile points in two different goal areas may not be the same, (i.e., the principal may not be indifferent between these two outcomes.)

The three words, "worth," "good," and "value" are often used synonymously with the term utility. However, in the last two paragraphs, these words have been used to express the decision maker's utility for only one goal area, and not the system as a whole. A key assumption of this model is that the principal's utility for a set of n percentile scores (one for each goal area, thus characterizing the state of the educational system) is simply the sum of his utility[3] for each of the individual scores;

[3]When these utilities are measured on appropriate scales.

therefore, it is evident that our next task is to transform the area scores into their associated "area utility values."

The above assumption about the decision maker's utility is equivalent to saying that his utility function is additively separable.[4] If we let

$n \equiv$ the number of terminal goal areas;

$a_i \equiv$ the (percentile) score obtained on a standardized test appropriate for measuring performance in area i;

$f_i(\cdot) \equiv$ the principals' standard[5] utility function for area i, i.e., it transforms the score for area i (a_i) into a number between 0 and 1 ($f_i(a_i)$);

$w_i \equiv$ the "weighting factor" for area i. If we require that $\sum_{i=1}^{n} w_i = 1$, this weight expresses the relative importance of area i with respect to the whole set of areas. The number w_i can also be viewed as the proper "scaling" factor for the standard utility function $f_i(\cdot)$ such that the principal's utility for the score a_i is given by the scaled utility function $\tilde{f}_i(a_i) = w_i f_i(a_i)$; and

$f(a_1,\ldots a_n) \equiv$ the principal's utility for the set of n scores a_i; i=1,---,n;

then the additive utility assumption says that:

$$(1)\quad f(a_1,a_2,\text{---},a_n) = \tilde{f}_1(a_1) + \tilde{f}_2(a_2) + \text{---} + \tilde{f}_n(a_n)$$

or equivalently:

$$(2)\quad f(a_1,a_2,\text{---},a_n) = w_1f_1(a_1) + w_2f_2(a_2) + \text{---} + w_nf_n(a_n)$$

---

[4] See Appendix II in Amor and Dyer (1970).

[5] The utility of a score of "0" is 0; the utility of a score of "100" is 1.

To transform the area score, a_i, into its associated area utility value, $\tilde{f}_i(a_i)$, we now need to determine the constant w_i and the function $f_i(\cdot)$

The values of the w_i's may be obtained in various ways. A procedure which illustrates one particular approach in terms of the 106 goals listed in the Appendix was described in the first paper of this symposium. This method allows the principal to gather information from several groups, including parents and teachers, regarding their priorities for student achievement in these 106 goal areas. Alternative methods are described in Fishburn (1967). The function $f_i(\cdot)$ can only be approximated through the analysis of empirical data.[6] However, a few statements can be made about its expected shape. It seems reasonable that a_i, defined in terms of a percentile score, will be considered to be of greater value as it increases; that is, if $a_i^1 > a_i^2$, then $f_i(a_i^1) > f_i(a_i^2)$. Consequently, we may assume that the function $f_i(\cdot)$ is monotonically increasing on the closed interval [0, 100], where the numbers in that interval refer to percentile scores. Additionally, there appear to be two particular percentile scores on a standardized test which serve as aspiration levels for the school principal: (1) the national norm (50th percentile score) and (2) the norms for schools of a particular "type," as characterized by the various environmental factors discussed earlier. Interviews with principals and the other individuals associated with elementary school systems indicated that an increase (in percentile score) of a given amount from a point below the national norm is considered to be of significantly greater value than the same amount of increase from a point above the

[6] For a discussion of several alternative approaches and a complete bibliography, see Fishburn (1967). For a discussion of how this was done in this particular project, see a forthcoming Center Report.

national norm. This suggests that the slope of f_i (·) is steeper at points below the 50th percentile score than at points above this score. A similar behavior may be "expected" with respect to the "environmental" norm; however, current data limitations prevent us from verifying this hypothesis. This prediction of a decreasing slope for the utility function as the percentile score increases is also consistent with the "law of diminishing marginal utility" which has been empirically verified in numerous studies in an economic context. One possible form of f_i (·) is shown in Fig. 1.

EXPECTED CHANGE IN UTILITY AS A DECISION CRITERION

This section describes how estimates of the decision maker's utility for various performance levels of the system may be combined with estimates of the effects on system performance of implementing various "programs" to provide a guide for decisions. In our particular application, the decision problem is to select educational goal areas which should receive more emphasis. The key assumption implicit in this discussion is that the decision maker prefers actions which maximize his expected utility.

It is reasonable to assume that a particular goal area will be selected for increased emphasis because (1) there exists an educational program (e.g., a new set of workbooks, the Sullivan reading program, a computer-assisted arithmetic program, etc.), which has a "reasonably good chance" of improving the student's performance in that area, and (2) a "significant" increase in the percentile score in that area will result in a "significant" increase in the decision maker's utility.

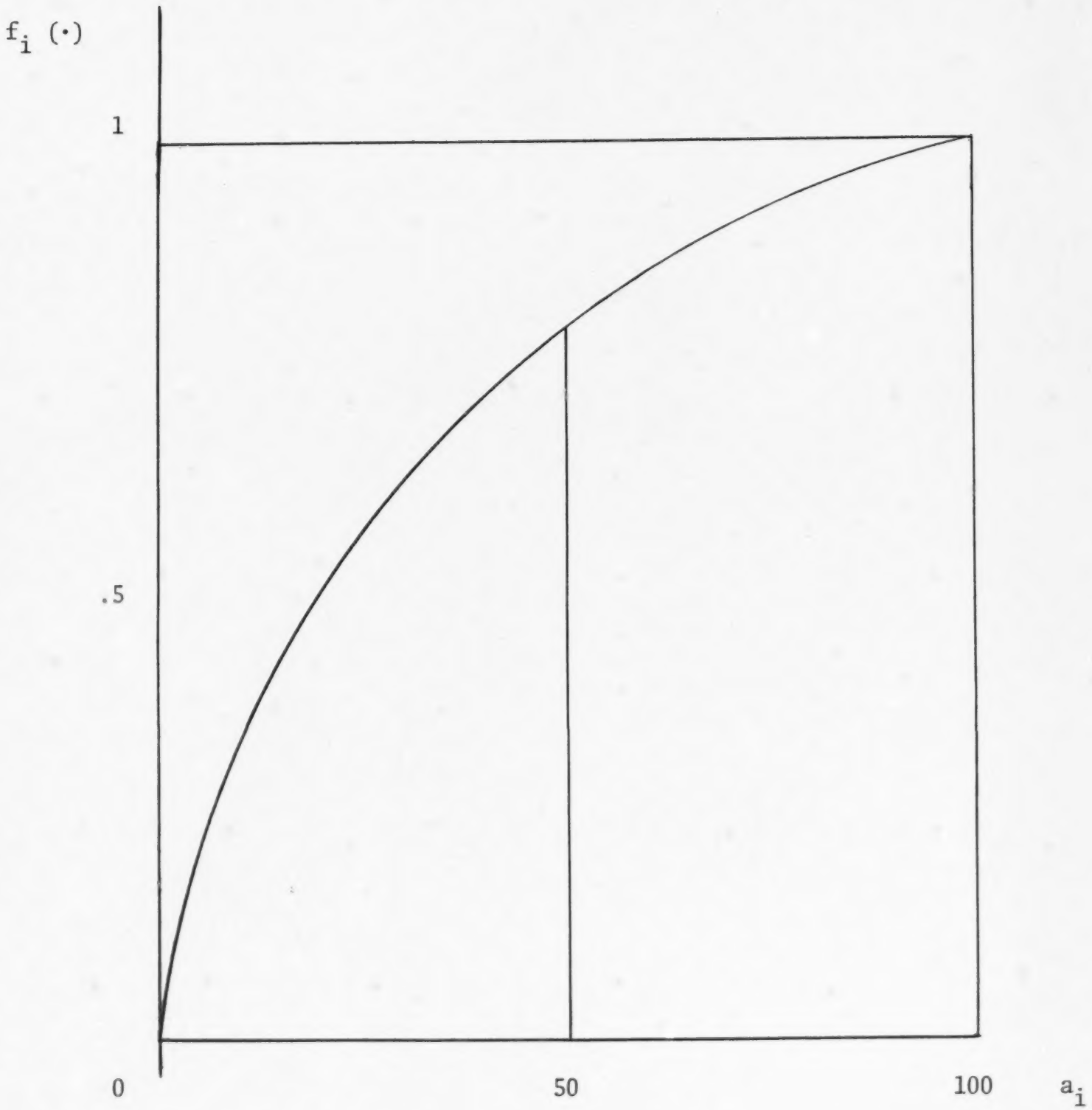

Figure 1 - A Possible Form of f_i (·)

Because of the interaction among the exogenous input factors, it is impossible to state that "the adoption of a program in area i will increase the performance from a_i° to $a_i^\circ + \delta_i$." Instead, the potential results of adopting a program in area i should be described by the conditional (or posterior) probability distribution of the scores which would be obtained upon retesting the students after implementation of the program. Using this distribution, one could calculate the probability of achieving a specified result. We assume that the random variable a'_i (representing the score to be obtained upon retesting) has a probability density function which depends on 2 factors: (1) the particular goal area (i), and (2) the current level of achievement (a_i°), and we denote this conditional probability distribution by $g_i\ (a'_i | a_i^\circ)$. For example, the goal area may be "operations with integers" and the current level of achievement may be "40th percentile ranking." A possible form of $g_i\ (a'_i\ |40)$ is shown in Figure 2.

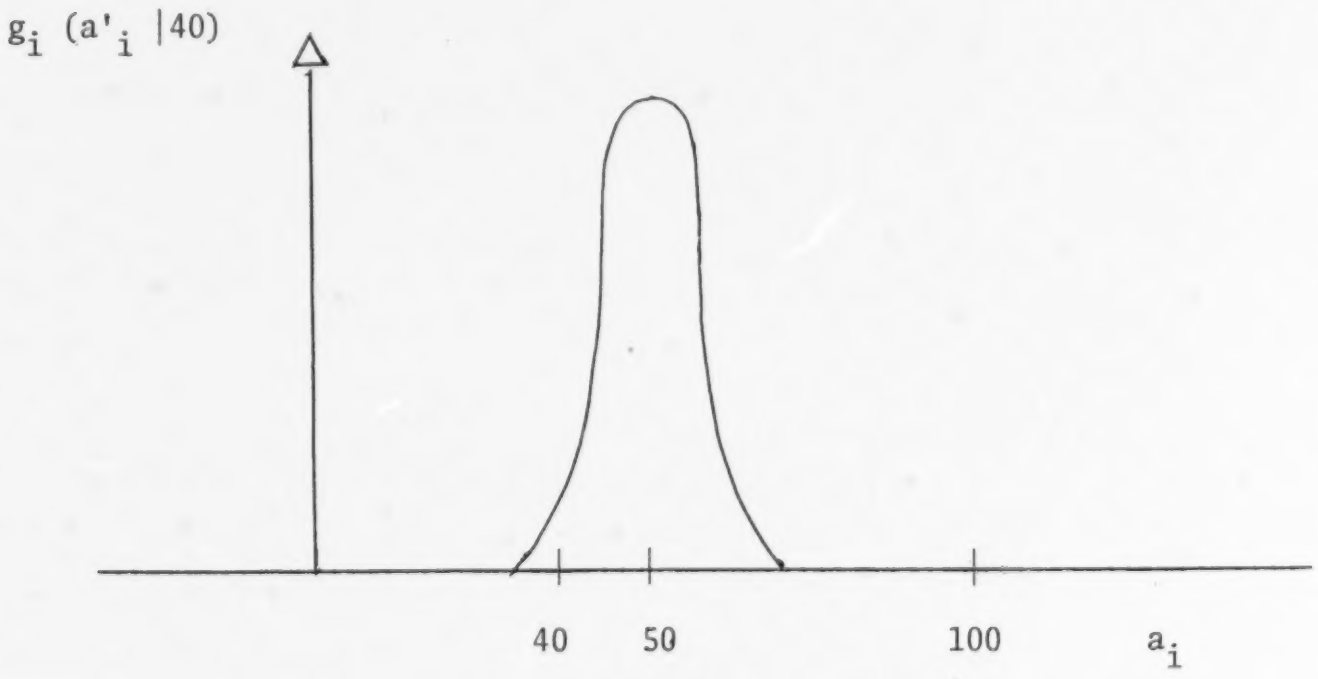

Figure 2
A Possible Form of a Posterior Probability Distribution of Scores

Unfortunately, it is extremely difficult and costly to obtain objective and generalizable information which would yield an estimate of this probability distribution. However, subjective information may be used to approximate the form of $g_i\ (a'_i\ |a_i^o)$. Persons who have observed the impact of new educational programs may be asked for their estimate of the probability of achieving a specific change in performance as a result of adopting a new program in area i, when the current level of performance is a_i^o. For example, these "experts" may estimate the probability of an increase of from three to five percentile points in area i, given a_i^o and the adoption of a new program, to be .3. This implies that

$$(3) \qquad \int_{a_i^o+3}^{a_i^o+5} g_i\ (a'_i\ |\ a_i^o)da'_i \approx .3.$$

We will denote a subjective estimate of the value of the integral of $g_i\ (a'_i\ |a_i^o)$ over a particular interval, Δ_k,[7] by $P_i\ (\Delta_k\ |\ a_i^o)$. By obtaining similar estimates of K non-overlapping intervals which cover the range of a'_i, we can obtain a discrete approximation to the desired distribution. By varying the size and number of the intervals we can make the approximation as "close" to the continuous function as we desire.[8] In addition, we naturally require that $\sum_{k=1}^{K} P_i\ (\Delta_k\ |\ a_i^o)=1$.

---

[7]Note that the symbol Δ_k carries information about both the location and the size of the interval of interest.

[8]See Schlaifer (1959) for a further discussion of this topic.

The theoretical probability distribution may be used to determine the expected change in utility, resulting from the implementation of a new program in area i given that the current level of performance is a_i, as follows:

$$(4)\quad E(\Delta U_i) = w_i\left[\int_0^{100} f_i(a'_i)g_i(a'_i \mid a_i^o)da'_i - f_i(a_i^o)\right] \quad 9$$

where E ≡ expected value operator,

ΔU_i ≡ change in utility in area i,

and the other variables are identical to those defined in the previous section. If the discrete approximation to $g_i(a'_i \mid a_i^o)$ is used, expression (4) simplifies to

$$(5)\quad E(\Delta U_i) = w_i\left[\{\sum_{k=1}^{K} f_i(\Delta_k)\cdot P_i(\Delta_k \mid a_i^o)\} - f_i(a_i^o)\right]$$

where the value of $f_i(\Delta_k)$ may be approximated by the value of $f_i(\cdot)$ at the midpoint of the interval Δ_k.

This number is the product of the measure of relative importance, w_i, the change in utility associated with a given change in performance, $f_i(\Delta_k) - f_i(a_i^o)$, and the subjective probability of that change, $P_i(\Delta_k \mid a_i^o)$. If the probabilities that the resulting score will fall in each of several disjoint intervals Δ_k are estimated, their product with their associated measures of change in utility must be summed. In words, expression (5) for the expected change in utility is approximately equivalent to

a measure of the importance of area i	·	the change in utility associated with a change in performance in area i	·	the probability of achieving that change

[9] See Appendix II of Amor and Dyer (1970) for a theoretical discussion of the assumptions implicit in this formulation.

IMPLEMENTING THE DECISION MODEL

This section will describe how the decision model could be used by an elementary school principal. This use would be encouraged by the provision of a series of tables containing the values computed from the expression in brackets in (5) for each area, i, and for a series of scores, a_i^o. The principal would merely be required to determine his own measures of the relative importance of each area (the w_i's), and obtain the remainder of the information directly from the appropriate table.

The computation of the "index numbers" or priority values for educational subject areas is quite simple and is outlined in a step-by-step procedure below. Figure 3 contains sample computations of priority values that are provided as an example.[10]

Step 1. The names of educational subject or "goal" areas for which priority values are desired are listed. This list should include all of the goal areas in which student performance has been assessed with a standardized testing instrument. The first column of Figure 3 shows ten goal areas for which priority values will be computed.

Step 2. For each of the goal areas listed in column 1, the current performance level (in school or class percentiles) is entered in column 2. For example, in column 2 of Figure 3 the current performance level for Creativity is the 43rd percentile.

---

[10]The authors wish to thank Dr. Paul Bradley for providing the example.

Step 3. The numerical value of the average rated importance of the goal area is entered in column 3. These values may be the averages obtained from a collective viewpoints rating of the goal areas as performed in Booklet II of the Elementary School Evaluation KIT described earlier, and correspond to the w_i's. In the example shown on Figure 3, the average rated importance of Creativity is 3.7, indicating that Creativity is a goal area which is moderately important.

Step 4. This is the first of the two steps that result in an entry for column 4, expected increase in utility. The first step is to select one of the several tables which are provided and in which the unscaled (standard) expected increases in utility (i.e., $\{\sum_{k=1}^{K} f_i(\Delta_k) \cdot P_i(\Delta_k \mid a_i^o)\} - f_i(a_i^o)$) are tabulated. This choice is made on the basis of typical level of student performance. The school type can be determined from a chart similar to the one presented in Figure 4.

Step 5. After the school type is determined, the corresponding table is used. That is, if a school is a type 1 school with typical achievement in the lower percentile range, then the principal should consult Table 1 to obtain estimates of unscaled expected increases in utility. The appropriate column in Table 1 is selected to correspond to the current level of student performance. If, for example, the current performance is at the 55th percentile, the sixth column of the table headed 51-60 would be used.

The expected unscaled increases in utility for the particular goal areas in question are found by searching down the appropriate columns. The computations in Figure 3 are based on a school whose typical level of student performance is low, so that the values in column 4 are taken from the figures presented in Table 1. Since the current level of student performance in Creativity is the 43rd percentile, the appropriate column in Table 1 is the

Figure 3

Sample computations of priority values

1	2	3	4	5	6
Goal Area	Current Performance Level (%ile)	Average Rated Importance	Expected Increase in Utility	Priority Value	Rank
Temperament: Social	57	3.2	017	54.4	10
Reasoning	21	3.3	035	115.5	3
Creativity	43	3.7	017	62.9	9
Language Construction	34	3.6	026	93.6	4
Arithmetic Operations	48	3.9	023	89.7	5
Health & Safety	22	2.5	032	80.0	6
Reading Comprehension	52	4.4	017	74.8	7
Scientific Knowledge	39	2.6	024	67.2	8
History & Civics	17	4.2	057	239.4	1
Sociology	28	4.1	032	131.2	2

Figure 4

Typical Levels of Student Performance

School Type	Characteristic
1	A school whose students are characteristically poor performers on standardized achievement tests. Many and/or most of the students' test scores fall in the bottom 25 to 30 per cent of a distribution based on a national sample. That is, the scores range from the 1st percentile to the 25th or 30th percentile.
2	A school whose students are characteristically average performers on standardized achievement tests. Many and/or most of the students' test scores fall in the middle 40 to 50 per cent of a distribution based on a national sample. That is, the scores range from the 25th to 30th percentile to the 70th or 75th percentile.
3	A school whose students are characteristically good performers on standardized achievement tests. Many and/or most of the students' test scores fall in the top 25 to 30 per cent of a distribution based on a national sample. That is, the scores range from the 70th or 75th percentile to the 99th percentile.

TABLE 1*

Expected increases in utility for school type 1

Goal Area	Current Performance Level in Percentiles								
	1-10	11-20	21-30	31-40	41-50	51-60	61-70	71-80	81-99
1. Temperament: Personal	091	056	036	025	020	017	014	011	010
2. Temperament: Social	091	056	036	025	020	017	014	011	010
3. Attitudes	091	056	036	025	020	017	014	011	010
4. Needs and Interest	091	056	036	025	020	017	014	011	010
5. Valuing Arts and Crafts	091	056	036	025	020	017	014	011	010
6. Producing Arts and Crafts	091	056	036	025	020	017	014	011	010
7. Understanding Arts and Crafts	091	056	036	025	020	017	014	011	010
8. Reasoning	061	055	035	025	020	016	014	012	010
9. Creativity	076	047	030	022	017	014	012	010	008
10. Memory	068	051	032	023	018	015	013	011	009
11. Foreign Language Skills	103	064	041	029	020	017	015	012	010
12. Foreign Language Assimilation	103	064	041	029	020	017	015	012	010
13. Language Construction	094	058	037	026	020	017	015	012	010
14. Reference Skills	094	058	037	026	020	017	015	012	010
15. Arithmetic Concepts	103	064	041	029	023	019	016	013	011
16. Arithmetic Operations	103	064	041	029	023	019	016	013	011
17. Mathematical Applications	103	064	041	029	023	019	016	013	011
18. Geometry	103	064	041	029	023	019	016	013	011
19. Measurement	103	064	041	029	023	019	016	013	011
20. Music Appreciation and Interest	091	056	036	025	020	017	014	011	010
21. Music Performance	091	056	036	025	020	017	014	011	010
22. Music Understanding	091	056	036	025	020	017	014	011	010
23. Health and Safety	081	050	032	023	018	015	013	010	008
24. Physical Skills	081	050	032	023	018	015	013	010	008
25. Sportsmanship	081	050	032	023	018	015	013	010	008
26. Physical Education	081	050	032	023	018	015	013	010	008
27. Oral-Aural Skills	094	058	037	026	021	017	015	012	010
28. Word Recognition	094	058	037	026	021	017	015	012	010
29. Reading Mechanics	094	058	037	026	021	017	015	012	010
30. Reading Comprehension	094	058	037	026	021	017	015	012	010
31. Reading Interpretation	094	058	037	026	021	017	015	012	010
32. Reading Appreciation and Response	094	058	037	026	021	017	015	012	010
33. Religious Knowledge	087	053	034	024	019	016	013	011	009
34. Religious Belief	087	053	034	024	019	016	013	011	009
35. Scientific Processes	087	054	034	024	019	016	013	011	009
36. Scientific Knowledge	087	054	034	024	019	016	013	011	009
37. Scientific Approach	087	054	034	024	019	016	013	011	009
38. History and Civics	092	057	036	026	020	017	014	012	010
39. Geography	087	053	034	024	019	016	013	011	009
40. Sociology	081	050	032	023	018	015	013	010	009
41. Application of Social Skills	087	053	034	024	019	016	013	011	009

*Based on hypothetical data

one with 41-50 at the top. The expected unscaled increase in utility is found by going down this column until the goal area Creativity is reached. This number is 017, and it is seen that this is the number which appears in column 4 of Figure 3.

Step 6. Now the priority value for a goal area can be computed. It is obtained by multiplying two numbers -- the rated importance and the unscaled expected increase in utility. These two numbers are found in columns 3 and 4 of Figure 3 respectively, and the product of the two numbers is entered in column 5. For the example of Creativity, its priority value is 3.7 X 017 = 62.9. This number corresponds to the results from expression (5) in the previous section.

Step 7. This last step is not performed until all the priority values have been computed. When this is accomplished, it is time to rank the goal areas on the basis of their priority values. The goal area with the highest priority value is given a rank of 1, the next highest is 2, and so on until all the goal areas have been ranked. In Figure 3 it is seen that the goal area of Creativity has a rank of 9, whereas the highest ranked goal area is History and Civics.

Now that the decision model has been used to compute priority values for some educational goal areas, and the educational goal areas have been ranked in terms of priority value, the next, and final, step is to implement the decision rule. The decision rule for the principal is elegantly simple: plan to revise the instructional program in the goal area that has the highest priority value. It is clear that some error could be incurred by the suggestion of this rule. Ideally, the principal should estimate his available resources, the resource requirements associated with each program, and solve the classical "knapsack

problem" by using these priority values in the objective function of an integer programming formulation. However, this process would require an expertise which cannot be assumed. Therefore, the simpler "rule of thumb" is suggested with certain caveats.

It is quite clear in the sample computations of Figure 3 that the priority value of History and Civics is easily the highest and that there is no other goal area which is close to History and Civics in priority value. This may not happen in all cases. For instance, if History and Civics were not included in Figure 3, then the highest ranked goal area would be Sociology, with a priority value of 131.2. Notice, however, that the priority value for Reasoning (the third ranked goal area) is 115.5, and the difference between it and Sociology is rather small compared to the difference in priority values for History and Civics and Sociology. When such a situation occurs, that is, when there are two or more goal areas with similar priority values, it may be best to suggest the temporary postponement of the decision to plan to revise the instructional program in a particular goal area. In lieu of making a final decision at this point, the principal should wait until a program evaluation is performed for each of the two or more goal areas that are similar in priority value. On the basis of these evaluations, he may then decide which one of the goal areas will receive a new instructional program. An alternative, and typical, solution would be to decide to plan revisions in the instructional programs of the two or three goal areas which are similar in priority value. This would be especially desirable if sufficient resources exist.

A further consideration in implementing the decision rule is the extent to which revising the instructional program in one goal area will

detract from achievement in other goal areas. There are no rules by which such deleterious side-effects can be determined, as they are unique to each school. Suffice it to say that the selection of one goal area does not imply that lesser efforts should be made in the remaining goal areas.

CONCLUSION

The model which was described in this paper provides the decision maker (an elementary school principal) with index numbers which represent estimates of the expected changes in his utility for the performance of his school which would result from the adoption of programs in particular areas. It is felt that these index numbers would provide valuable information to the decision maker for dealing with the problem of identifying areas in which action should be considered, and identifying the types of programs which would provide the greatest expected contribution to the achievement of his instructional goals. The model provides the basis for Booklet V of the Elementary School Evaluation Kit: Needs Assessment.

REFERENCES

1. Amor, J. P., & Dyer, J. S. A Decision Model for Evaluating Potential Change in Instructional Programs. CSE Report No. 62. Los Angeles: Center for the Study of Evaluation, University of California, 1970.

2. Hoepfner, R., Klein, S. P., & Bradley, P. A. Elementary school evaluation KIT: Needs assessment. Los Angeles: Center for the Study of Evaluation, UCLA, 1970.

3. Hoepfner, R., Strickland, G., Stangel, G., Jansen, P., & Patalino, M. CSE elementary school test evaluations. Los Angeles: Center for the Study of Evaluation, 1970.

4. Fishburn, P. C. Methods of estimating additive utilities. Management Science, 1967, 13 (7).

5. Schlaifer, R. Probability and statistics for business decisions. New York: McGraw-Hill, 1959.

6. Von Neumann, J., & Morgenstern, O. Theory of games and economic behavior. Princeton, N. J.: Princeton, University Press, 1953.

APPENDIX

OUTLINE OF 145 GOALS of Elementary School Education from the CSE Elementary School Evaluation KIT

OUTLINE OF 145 GOALS

OF ELEMENTARY SCHOOL EDUCATION

AFFECTIVE

1. TEMPERAMENT: PERSONAL
   - A. Shyness-Boldness
   - B. Neuroticism-Adjustment
   - C. General Activity-Lethargy

2. TEMPERAMENT: SOCIAL
   - A. Dependence-Independence
   - B. Hostility-Friendliness
   - C. Socialization-Rebelliousness

3. ATTITUDES
   - A. School Orientation
   - B. Self Esteem

4. NEEDS AND INTERESTS
   - A. Need Achievement
   - B. Interest Areas

ARTS-CRAFTS

5. VALUING ARTS AND CRAFTS
   - A. Appreciation of Arts and Crafts
   - B. Involvement in Arts and Crafts

6. PRODUCING ARTS AND CRAFTS
   - A. Representational Skill in Arts and Crafts
   - B. Expressive Skill in Arts and Crafts

7. UNDERSTANDING ARTS AND CRAFTS
   - A. Arts and Crafts Comprehension
   - B. Developmental Understanding of Arts and Crafts

COGNITIVE

8. REASONING
   - A. Classificatory Reasoning
   - B. Relational-Implicational Reasoning
   - C. Systematic Reasoning
   - D. Spatial Reasoning

9. CREATIVITY
   - A. Creative Flexibility
   - B. Creative Fluency

10. MEMORY
    A. Span and Serial Memory
    B. Meaningful Memory
    C. Spatial Memory

FOREIGN LANGUAGE

11. FOREIGN LANGUAGE SKILLS
    A. Reading Comprehension of a Foreign Language
    B. Oral Comprehension of a Foreign Language
    C. Speaking Fluency in a Foreign Language
    D. Writing Fluency in a Foreign Language

12. FOREIGN LANGUAGE ASSIMILATION
    A. Cultural Insight through a Foreign Language
    B. Interest in and Application of a Foreign Language

LANGUAGE ARTS

13. LANGUAGE CONSTRUCTION
    A. Spelling
    B. Punctuation
    C. Capitalization
    D. Grammar and Usage
    E. Penmanship
    F. Written Expression
    G. Independent Application of Writing Skills

14. REFERENCE SKILLS
    A. Use of Data Sources as Reference Skills
    B. Summarizing Information for Reference

MATHEMATICS

15. ARITHMETIC CONCEPTS
    A. Comprehension of Numbers and Sets in Mathematics
    B. Comprehension of Positional Notation in Mathematics
    C. Comprehension of Equations and Inequalities
    D. Comprehension of Number Principles

16. ARITHMETIC OPERATIONS
    A. Operations with Integers
    B. Operations with Fractions
    C. Operations with Decimals and Percents

17. MATHEMATICAL APPLICATIONS
    A. Mathematical Problem Solving
    B. Independent Application of Mathematical Skills

18. GEOMETRY
    A. Geometric Facility
    B. Geometric Vocabulary

19. MEASUREMENT
    A. Measurement Reading and Making
    B. Statistics

MUSIC

20. MUSIC APPRECIATION AND INTEREST
    A. Music Appreciation
    B. Music Interest and Enjoyment

21. MUSIC PERFORMANCE
    A. Singing
    B. Musical Instrument Playing
    C. Dance (Rhythmic Response)

22. MUSIC UNDERSTANDING
    A. Aural Identification of Music
    B. Music Knowledge

PHYSICAL EDUCATION - HEALTH - SAFETY

23. HEALTH AND SAFETY
    A. Practicing Health and Safety Principles
    B. Understanding Health and Safety Principles
    C. Sex Education

24. PHYSICAL SKILLS
    A. Muscle Control (Physical Education)
    B. Physical Development and Well-Being (Physical Education)

25. SPORTSMANSHIP
    A. Group Activity - Sportsmanship
    B. Interest in and Independent Participation in Sports and Games

26. PHYSICAL EDUCATION
    A. Understanding of Rules and Strategies of Sports and Games
    B. Knowledge of Physical Education Apparatus and Equipment

READING

27. ORAL-AURAL SKILLS
    A. Listening Reaction and Response
    B. Speaking

28. WORD RECOGNITION
    A. Phonetic Recognition
    B. Structural Recognition

29. READING MECHANICS
    A. Oral Reading
    B. Silent Reading Efficiency

30. READING COMPREHENSION
    - A. Recognition of Word Meanings
    - B. Understanding Ideational Complexes
    - C. Remembering Information Read

31. READING INTERPRETATION
    - A. Inference Making from Reading Selections
    - B. Recognition of Literary Devices
    - C. Critical Reading

32. READING APPRECIATION AND RESPONSE
    - A. Attitude toward Reading
    - B. Attitude and Behavior Modification from Reading
    - C. Familiarity with Standard Children's Literature

RELIGION

33. RELIGIOUS KNOWLEDGE

34. RELIGIOUS BELIEF

SCIENCE

35. SCIENTIFIC PROCESSES
    - A. Observation and Description in Science
    - B. Use of Numbers and Measures in Science
    - C. Classification and Generalization in Science
    - D. Hypothesis Formation in Science
    - E. Operational Definitions in Science
    - F. Experimentation in Science
    - G. Formulation of Generalized Conclusions in Science

36. SCIENTIFIC KNOWLEDGE
    - A. Knowledge of Scientific Facts and Terminology
    - B. The Nature and Purpose of Science

37. SCIENTIFIC APPROACH
    - A. Science Interest and Appreciation
    - B. Application of Scientific Methods to Everyday Life

SOCIAL STUDIES

38. HISTORY AND CIVICS
    - A. Knowledge of History
    - B. Knowledge of Governments

39. GEOGRAPHY
    - A. Knowledge of Physical Geography
    - B. Knowledge of Socio-Economic Geography

40. SOCIOLOGY
    - A. Cultural Knowledge
    - B. Social Organization Knowledge

41. APPLICATION OF SOCIAL STUDIES
    A. Research Skills in Social Studies
    B. Citizenship
    C. Interest in Social Studies

This publication is published pursuant to a contract with the U.S. Office of Education, Department of Health, Education and Welfare. Points of view or opinions stated do not necessarily represent official U.S.O.E. position or policy.